Chapter 5 Markov Chain Monte Carlo (MCMC)

- powerful sampling algorithms, associated with Bayesian statistics

Given a probability distribution π, the goal of MCMC methods is to simulate a random variable X whose distribution is π.

The idea is to construct an appropriate ergodic MC with limiting distr. π. Run that Markov chain long enough to converge to that long-term π ("burn-in"), then start using the samples generated by the chain.

Recall Law of Large Numbers: given iid sequence $Y_1, Y_2, Y_3, \cdots$ of RVs with common mean $\mu < \infty$, we have $\lim_{n\to\infty} \frac{1}{n}(Y_1 + \cdots + Y_n) = \mu$ with prob 1.

(That is, the sample mean approaches the true mean as sample size $\to \infty$.)

Also works for bounded functions: $\lim_{n\to\infty} \frac{f(Y_1) + \cdots + f(Y_n)}{n} = \mathbb{E}[f(Y)]$

Theorem 5.1 Strong law of large numbers for Markov chains

Suppose $X_0, X_1, \cdots$ is an ergodic MC with stationary distr $\bar{\pi}$ and state space S.

ergodic = irreducible, aperiodic, all states have finite expected return times

Let f be a bounded, real-valued function and X an RV with distr $\bar{\pi}$.

Then with prob 1, $$\lim_{n\to\infty} \frac{f(X_0)+\cdots+f(X_{n-1})}{n} = \mathbb{E}[f(X)]$$

$$= \sum_{j\in S} f(j)\pi_j$$

Note: RV X is distinct from RVs in chain $X_0, X_1, \cdots$, but shares same $\bar{\pi}$.

Markov chains are not independent sequences (X_{n+1} depends on X_n), but it is true that, for ergodic chains, successive excursions between visits to the same state are independent (like restarting chain).

Given an ergodic Markov chain $X_0, X_1, \cdots$ with stationary distribution $\bar{\pi}$, let A be a nonempty set of states and set $\pi_A = \sum_{j \in A} \pi_j$ (will show this is the long-term expected proportion of visits to set A)

Let $I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$ (indicator fn for set A)

Then $\sum_{k=0}^{n-1} I_A(X_k) = \#$ visits to states in set A in steps 0 to $n-1$.

According to Thm 5.1, with prob 1, $\lim_{n \to \infty} \frac{I_A(X_0) + \cdots + I_A(X_{n-1})}{n}$

$$= \mathbb{E}[I_A(X)] = \mathbb{P}\{X \in A\}$$

$$= \pi_A$$

So the long-term expected proportion of visits to states in set A is π_A.

Example: binary sequence with no adjacent 1s

00101000101000010 ...

We want an algorithm that samples a sequence of length m of 0s and 1s with no adjacent 1s (call this a "good" sequence).

Challenge: "good" sequences are rare in the set of binary sequences.

$m=4$: $2^4 = 16$ possible binary sequences, with 8 "good" seq.

$m=100$: $2^{100} \approx 10^{30}$ possible binary seq, with 10^{21} "good" seq

(prob 10^{-9} of randomly choosing a good seq)

Suppose you want to efficiently generate a randomly chosen good seq.

1) Brute force: store all possible good sequences, randomly choose one.

2) Rejection method: generate random binary sequences through m coin flips, reject if not good seq, repeat until obtain good seq.

3) MCMC approach: construct ergodic Markov chain $X_0, X_1, \cdots$ whose state space is the set of good sequences and whose limiting distribution is uniform.

↖ don't need to store, unlike brute force approach

MCMC approach: construct ergodic Markov chain $X_0, X_1, \cdots$
whose state space is the set of good sequences and
whose limiting distribution is uniform.

Algorithm: Start with a good sequence as X_0 state (of length m)
Pick one of m components at random
If component is 1, flip to 0 (still a good seq)
If component is 0, switch to 1 if results in
a good seq, otherwise leave as 0.

R demo

The underlying idea here is that of the Metropolis-Hastings algorithm.

Section 5.2 Metropolis-Hastings Algorithm

↖ developed in 1950s ↖ extended in 1970

Given a discrete prob distribution $\bar{\pi}$ ("target distr"), we want to construct a reversible Markov chain $X_0, X_1, \cdots$ whose stationary distribution is $\bar{\pi}$.

(Reversible if $\pi_i P_{ij} = \pi_j P_{ji}$ for all states i & j.

Time-reversal chain is $\tilde{P}_{ij} = \frac{\pi_j P_{ji}}{\pi_i}$.)

Let T be any transition matrix on the same state space as $X_0, X_1, \cdots$,

Choosing T to represent a Markov chain that is easy to work with.

It will be used to generate "proposal states" (like flipping a random bit in the binary seq example).

Algorithm then decides whether or not to accept the proposal state (or stay at the current state).

Metropolis-Hastings Algorithm

Suppose we're at the n^{th} step of the chain with $X_n = i$.

① Choose a state j from the state space with probability T_{ij}

proposal state — proposal matrix

② Decide whether to accept the proposal state:

Let $a(i,j) = \dfrac{\pi_j T_{ji}}{\pi_i T_{ij}}$ acceptance function

If $a(i,j) \geq 1$, accept the proposal state j

Otherwise accept j with probability $a(i,j)$ (keep i if reject j)

That is, $X_{n+1} = \begin{cases} j & \text{if } U \leq a(i,j) \\ i & \text{if } U > a(i,j) \end{cases}$ where $U \sim \text{Unif}(0,1)$

Theorem The sequence $X_0, X_1, \cdots$ generated by the MH algorithm is a reversible Markov chain with stationary distribution $\vec{\pi}$.

Proof: Markov because X_{n+1} only depends on X_n.

Let P be the transition matrix resulting from the MH algorithm applied to the proposal matrix T.

We need to prove $\pi_i P_{ij} = \pi_j P_{ji}$ for all states i & j, $i \neq j$.

Prob of j being the proposal state is T_{ij}

Prob of j being accepted is $\min(a(i,j), 1)$

$$\mathbb{P}\{U \leq a(i,j)\} = \begin{cases} a(i,j) & \text{if } a(i,j) \leq 1 \\ 1 & \text{if } a(i,j) > 1 \end{cases}$$

uniform RV on $[0,1]$

$$\text{So } P_{ij} = \begin{cases} T_{ij}\, a(i,j) & \text{if } \pi_j T_{ji} \leq \pi_i T_{ij} \\ T_{ij} & \text{if } \pi_j T_{ji} > \pi_i T_{ij} \end{cases} \qquad = P_{ji}$$

Case ①: If $\pi_j T_{ji} \leq \pi_i T_{ij}$. then $\pi_i P_{ij} = \pi_i T_{ij} \cdot \frac{\pi_j T_{ji}}{\pi_i T_{ij}} = \pi_j T_{ji}$

Case ②: If $\pi_j T_{ji} > \pi_i T_{ij}$, then $\pi_i P_{ij} = \pi_i T_{ij} \cdot \frac{\pi_j T_{ji}}{\pi_j T_{ji}} = \pi_j T_{ji}\, a(j,i)$

$= \pi_j P_{ji}$ because $a(j,i) \leq 1$

So MH chain is reversible, with stationary distr. $\vec{\pi}$.

Numerical issues with MCMC

① Burn-in: need to run chains long enough that prob distr. of visiting states has converged to target distr. π (discard first part, only keep later samples)

② Thinning: successive steps will be correlated (X_{n+1} depends on X_n) So run many chains, discard burn-in portion, then keep every k^{th} value (k depends on how long correlation takes to decay).

③ Mixing: check that not too many proposal states are rejected (could indicate poor choice of proposal transition matrix T)

MCMC software typically provides diagnostics to check these and other issues, to ensure good results.

Why MCMC works well for Bayesian methods

Let D be the observed data and θ represent model parameters

Posterior distribution

$$\underbrace{P\{\theta \mid D\}}_{\text{target distr}} = \frac{P\{\theta\} P\{D \mid \theta\}}{P\{D\}}$$

π

prior — likelihood of the observed data being generated by the model

$P\{D\}$: prob of these data being observed - usually unknown

MH comes to the rescue because only the ratio $\frac{\pi_j}{\pi_i}$ is needed, canceling out the unknown $P\{D\}$.

R demo - decryption lab

Section 5.3 Gibbs sampler

Aimed at high-dimensional state spaces

Idea is to change one dimension at a time, conditioned on other dim.

Suppose the problem involves m variables $x_1, \cdots, x_m$

The target distribution in an m-dimensional joint density $\pi(\bar{x}) = \pi(x_1, \cdots, x_m)$

Bivariate example

Univariate normal density $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$ mean μ, variance σ^2

Let Z_1 and Z_2 be independent standard normal RVs ($\mu = 0$, $\sigma = 1$)

Let $X_1 = Z_1 + \mu_1$

$X_2 = \rho Z_1 + \sqrt{1-\rho^2}\, Z_2 + \mu_2$

That is, $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$

Covariance $\Sigma = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} 1 & \rho \\ 0 & \sqrt{1-\rho^2} \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$

Joint density $f(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left((x_1-\mu_1)^2 - 2\rho(x_1-\mu_1)(x_2-\mu_2) + (x_2-\mu_2)^2\right)}$

Conditional distr. $f(x_1 | x_2) = \frac{f(x_1, x_2)}{\int_{-\infty}^{\infty} f(y, x_2)\, dy} = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{(x_1 - \mu_1 - \rho(x_2-\mu_2))^2}{1-\rho^2}}$

→ univariate normal RV with mean $\mu_1 + \rho(x_2 - \mu_2)$ & variance $1-\rho^2$

R demo